DISCOVER

www.discover-earth.org

Service Oriented Architecture for Harvesting Distributed Data Repositories

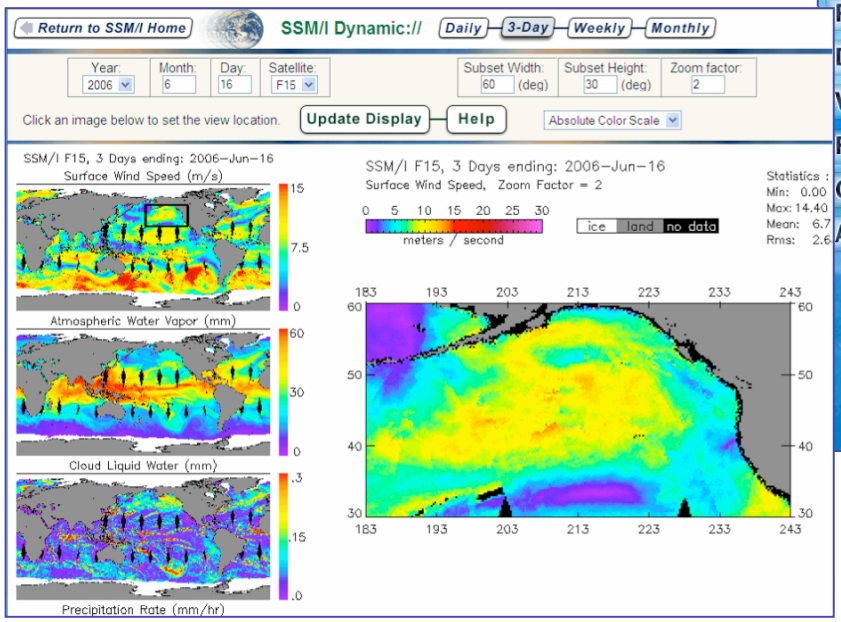
Helen Conover, Sara Graves, Ken Keiser, Lamar Hawkins
Information Technology &Systems Center
University of Alabama in Huntsville
Huntsville, AL 35899
www.itsc.uah.edu

Presented by Danny Hardin

NASA Earth Science Technology Conference
College Park, MD
27-29 June 2006

DISCOVER

The primary objective of the DISCOVER Project is to provide highly accurate, long-term ocean and climate products suitable for the most demanding Earth research applications via easy-to-use display and data access tools.



Remote Sensing Systems
www.remss.com

DISCOVER
www.discover-earth.org

Relevance of DISCOVER Technology

- **Original REASoN solicitation**
  - Apply principles from the Strategy for the Evolution of ESE Data Systems (SEEDS) regarding **standards and interfaces for interoperability** and exchange of data and information
  - Develop applicable **advanced data systems technologies** integrated into the project, including
    - Data and service locator technologies **leveraging the web and commercial approaches** that are tailored to the unique demands of the geo-spatial Earth science data sets
    - Techniques addressing **seamless, automated access** to data residing in **multiple distributed archives**
    - **Languages and metadata techniques** enabling Earth science community-centric services and open tool sets
- **NASA's Vision for Evolution of EOSDIS Elements**
  - Seamlessly combining **multiple data and metadata streams**
  - **Distributed repositories** – physical location of data irrelevant
  - Service oriented architecture with **modular components** and **machine-to-machine interfaces**
  - **Custom processing** to provide only the data needed, the way it is needed
  - **Open interfaces** and **standard protocols** for interoperability with other relevant systems

DISCOVER Information Technology Objectives

Develop an open, distributed, heterogeneous data system designed to:

- Improve access and use of data and information products
- Extend online data distribution
- Provide for automated data management and distribution
- Provide automated order tracking and metrics

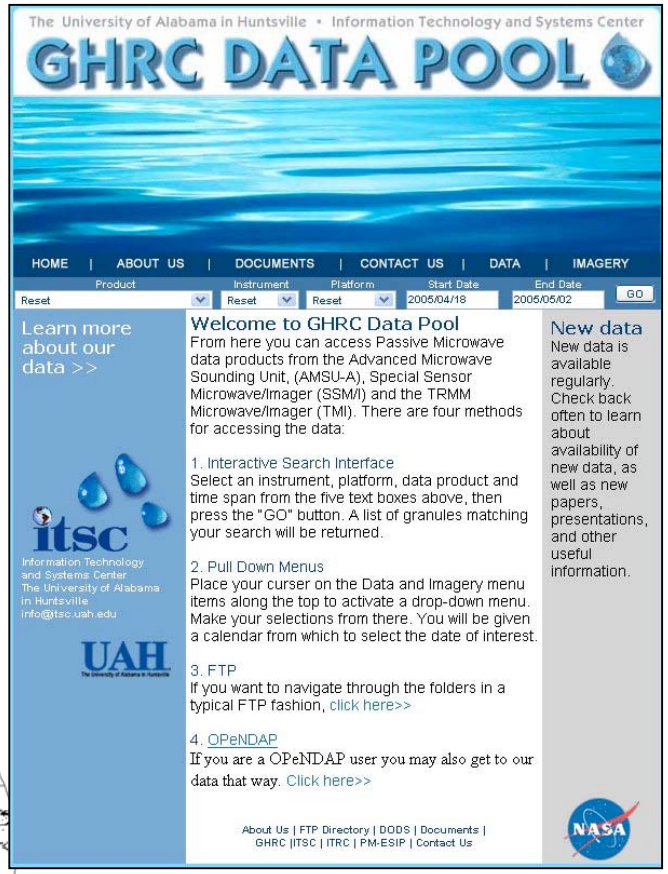
DISCOVER Information Technology Objectives

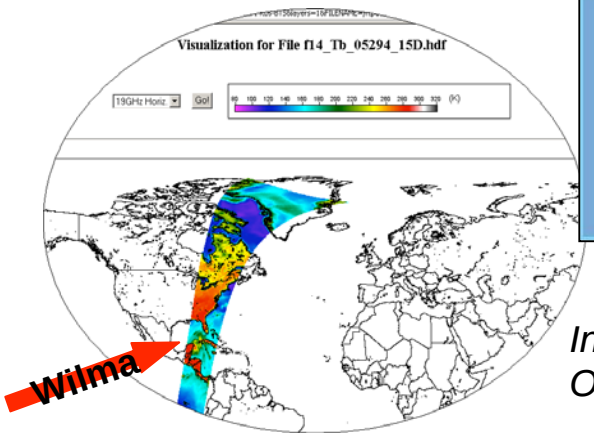
Develop a distributed, service oriented architecture for managing multiple distributed data repositories with modular components:

- Central Metadata Catalog Services
- Search and Order Interfaces
- Order Tracking Services
- Subsetting
- Visualization Through Web Mapping Services
- Order Broker
- Packaging Services

GHRC Data Pool

The GHRC Data Pool provides a data search-and-order interface for data in distributed repositories maintained by a single management infrastructure. The next generation of this data pool is integrating data from repositories maintained by different institutions so it has been necessary to develop tools and functionality capable of **consistently collecting and managing metadata across the distributed online resources.**





Interactive browse images via OGC web mapping services

DISCOVER
www.discover-earth.org

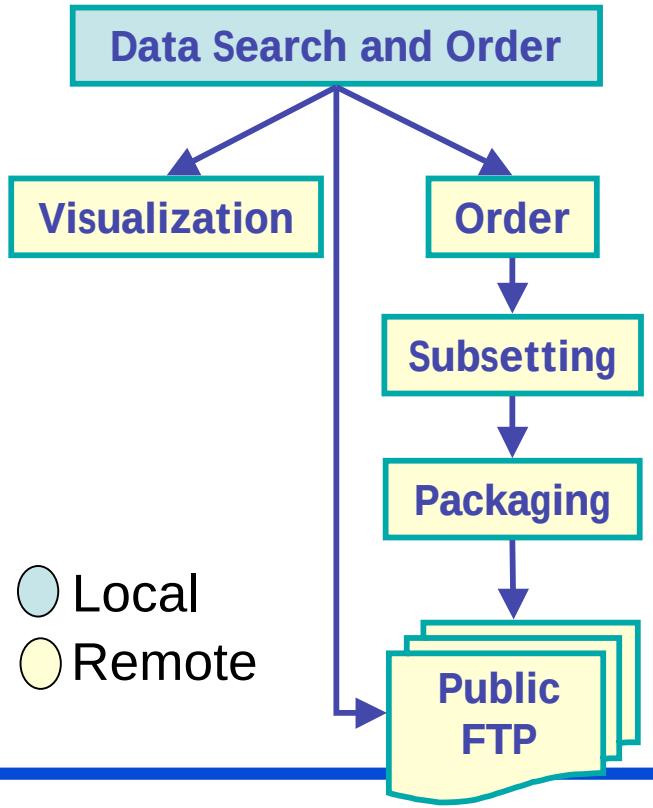
Remote Sensing Systems
www.remss.com

NASA

itsc UAH

The University of Alabama in Huntsville • Information Technology and Systems Center

GHRC DATA POOL



- On-line data access with integrated data services
- Automated ordering, subsetting, packaging, display and delivery of scientific data
- Multiple distributed repositories at UAH and RSS
- Common user interface, **data catalog** and order tracking

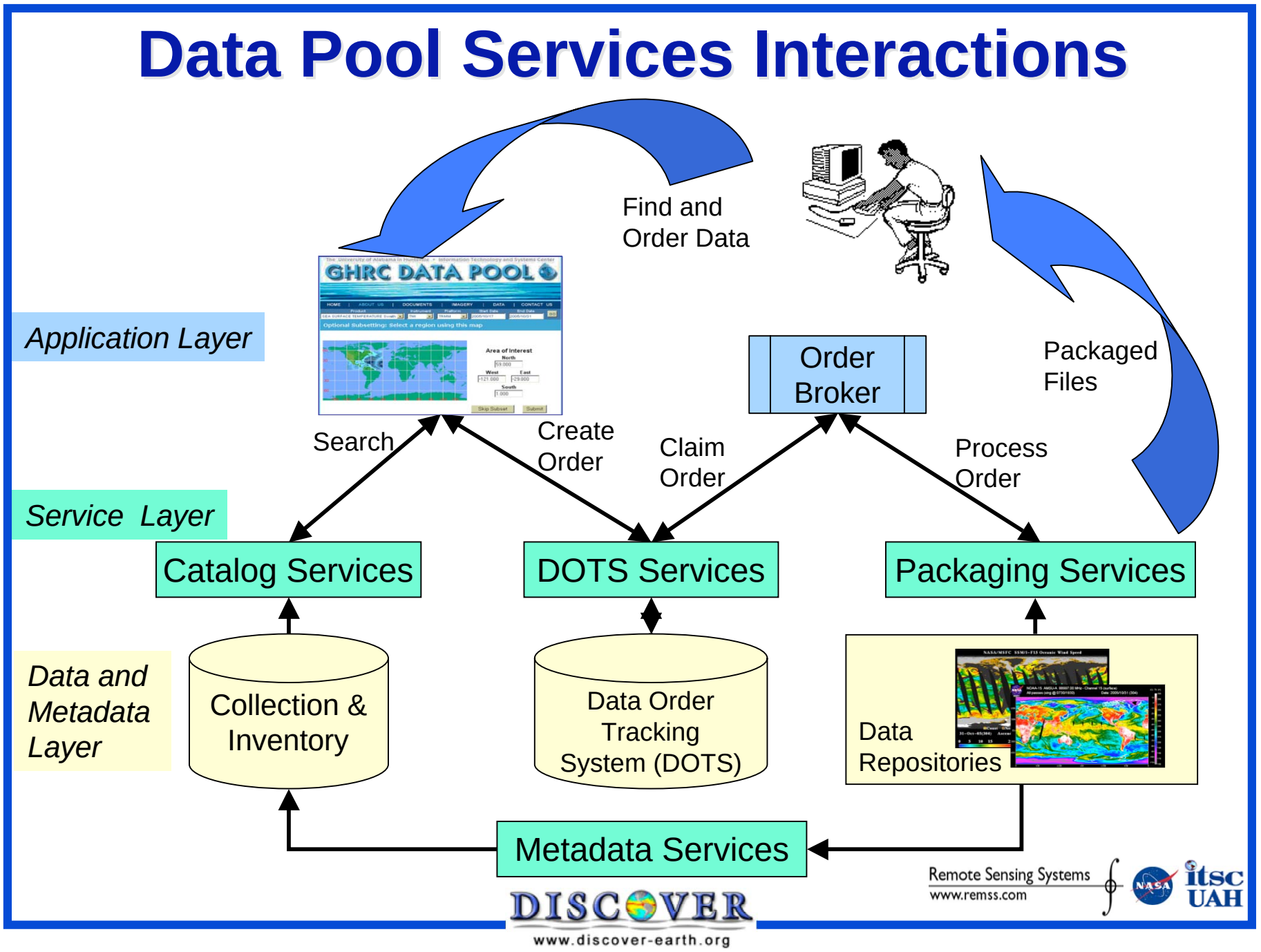
http://datapool.nsstc.nasa.gov

DISCOVER
www.discover-earth.org

Remote Sensing Systems
www.remss.com

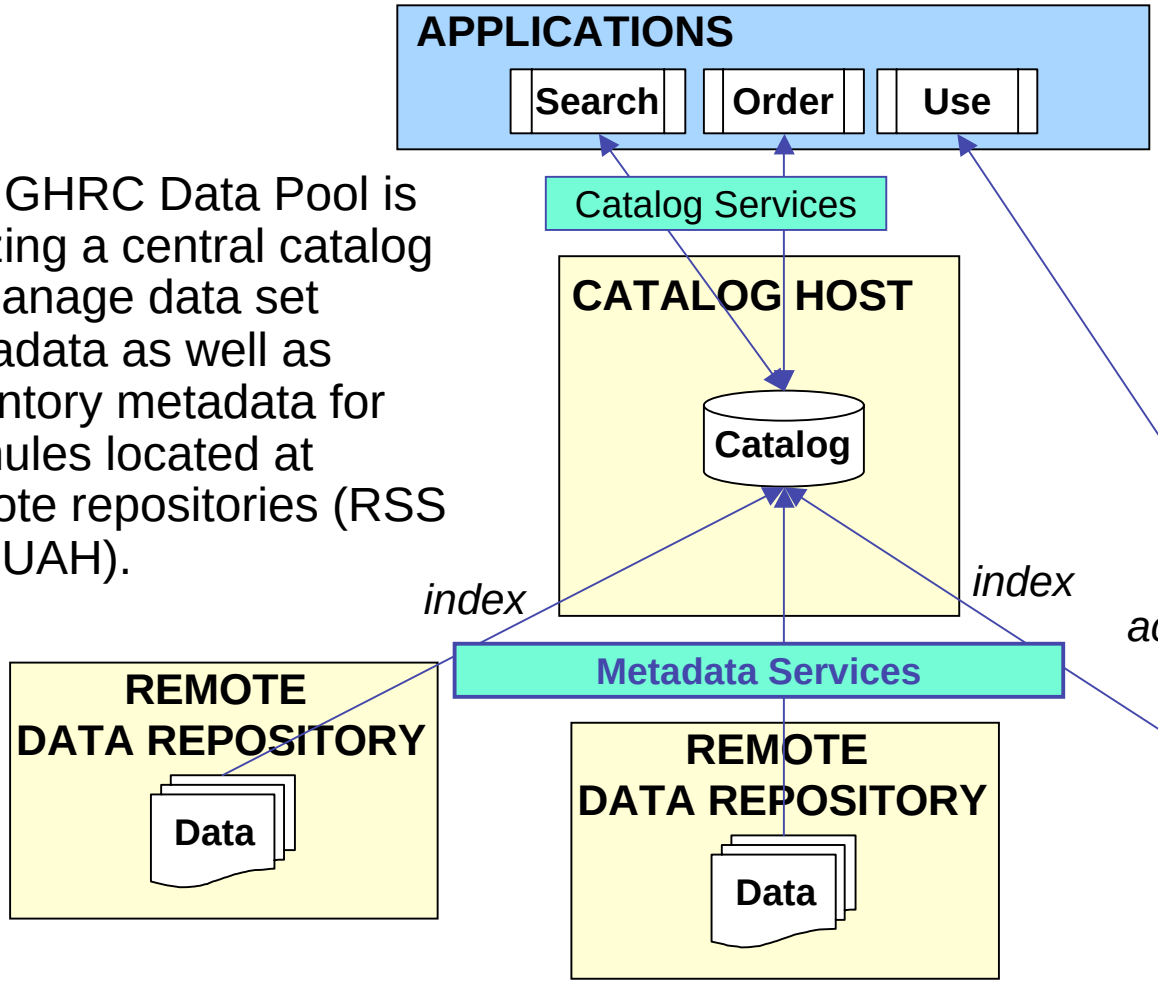
itsc UAH

Data Pool Services Interactions



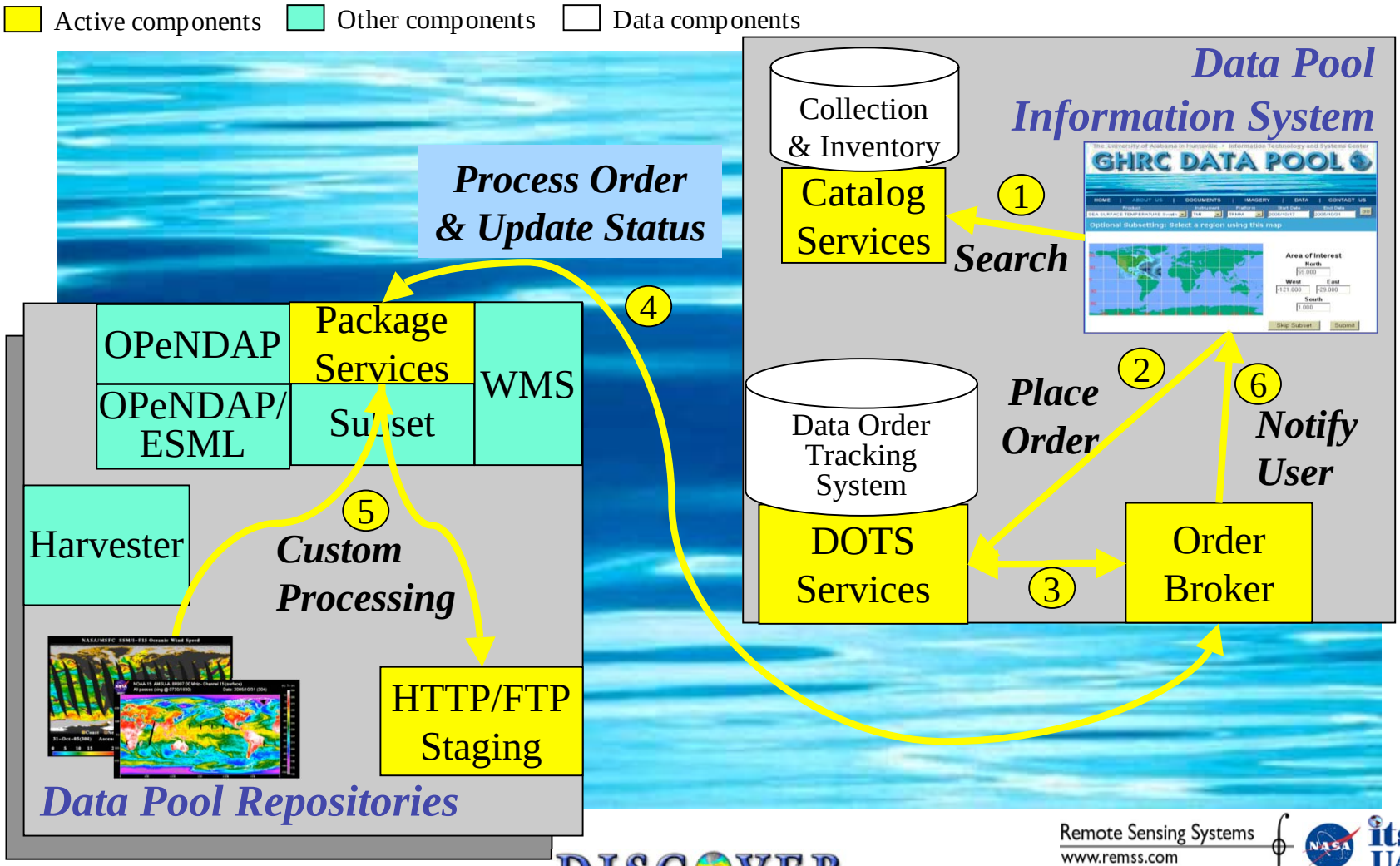
Data Pool Approach: Distributed Data Repositories

The GHRC Data Pool is utilizing a central catalog to manage data set metadata as well as inventory metadata for granules located at remote repositories (RSS and UAH).

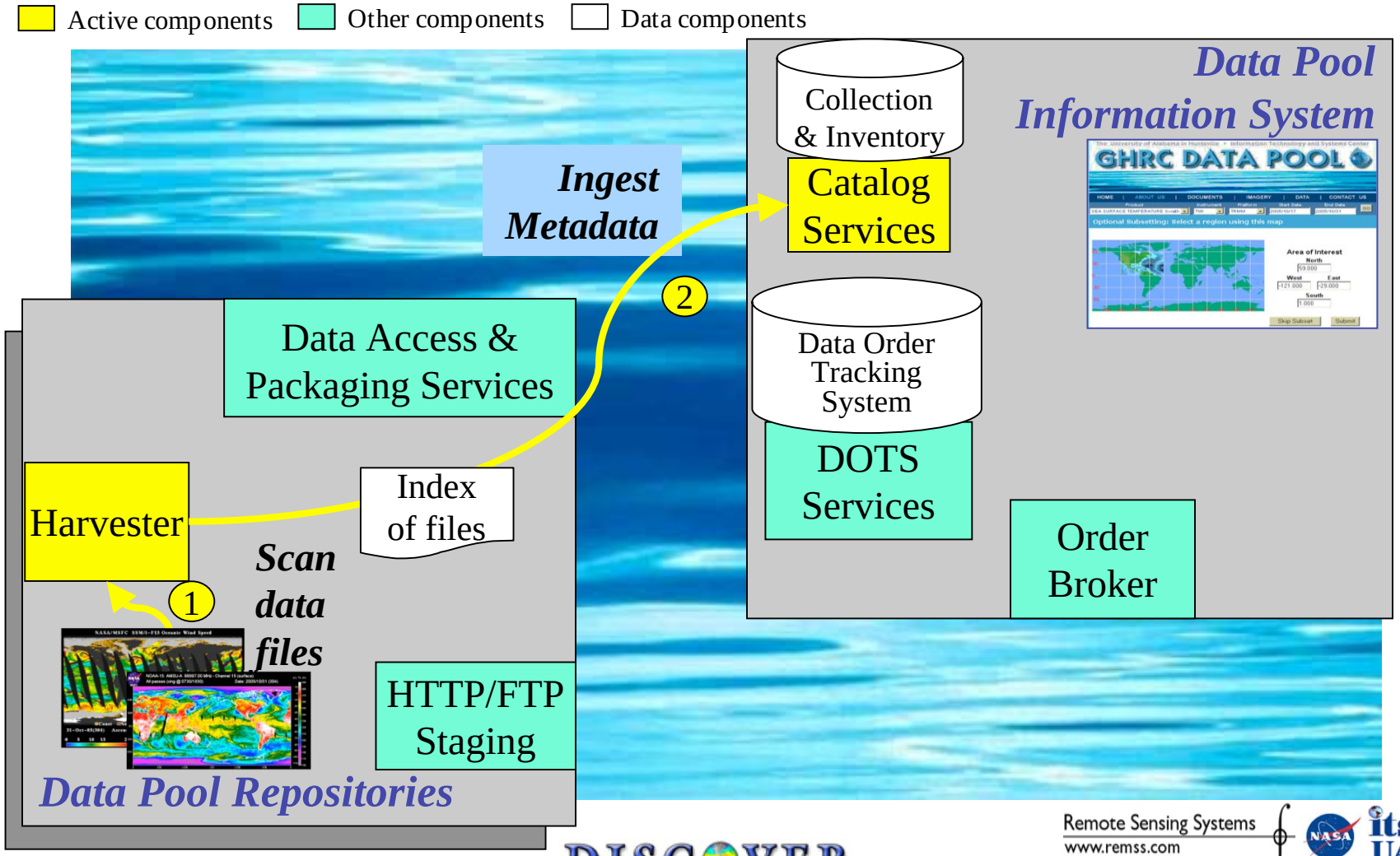


Web service interfaces support the distributed nature of the system. This helps avoid direct access to the database which can introduce security concerns.

Data Pool Services Use Case:
Automated Order Processing



Data Pool Services Use Case:
Automated Metadata Ingest



DISCOVER
www.discover-earth.org

Remote Sensing Systems
www.remss.com

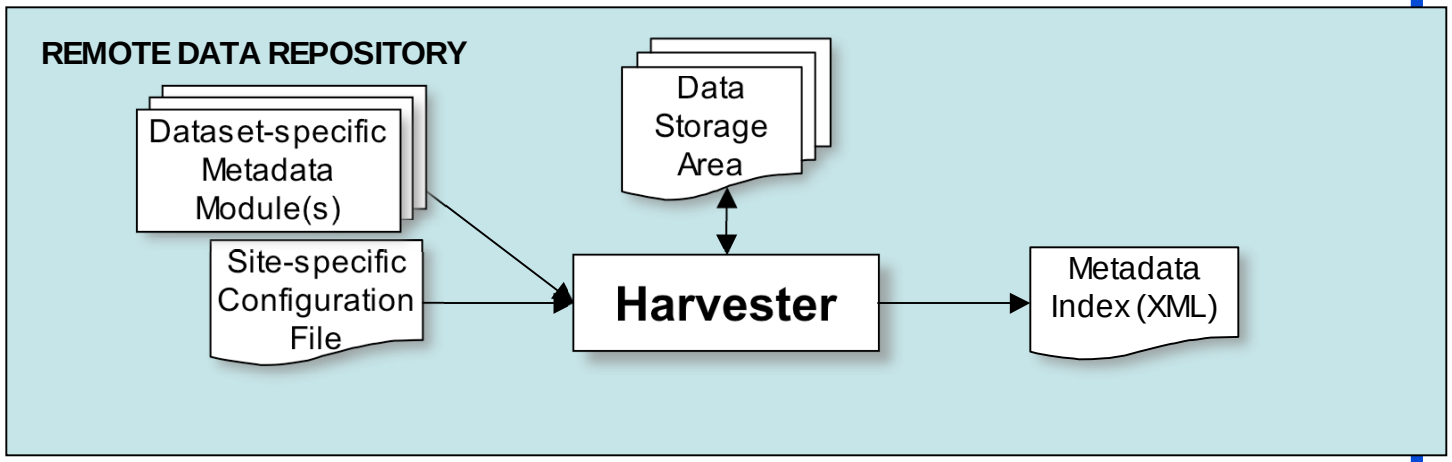
NASA itsc UAH

Harvester

The goal is gathering metadata for data on remote data repositories.

The Harvester is deployed on the remote repository, requiring minor site-specific configurations and dataset-specific metadata modules.

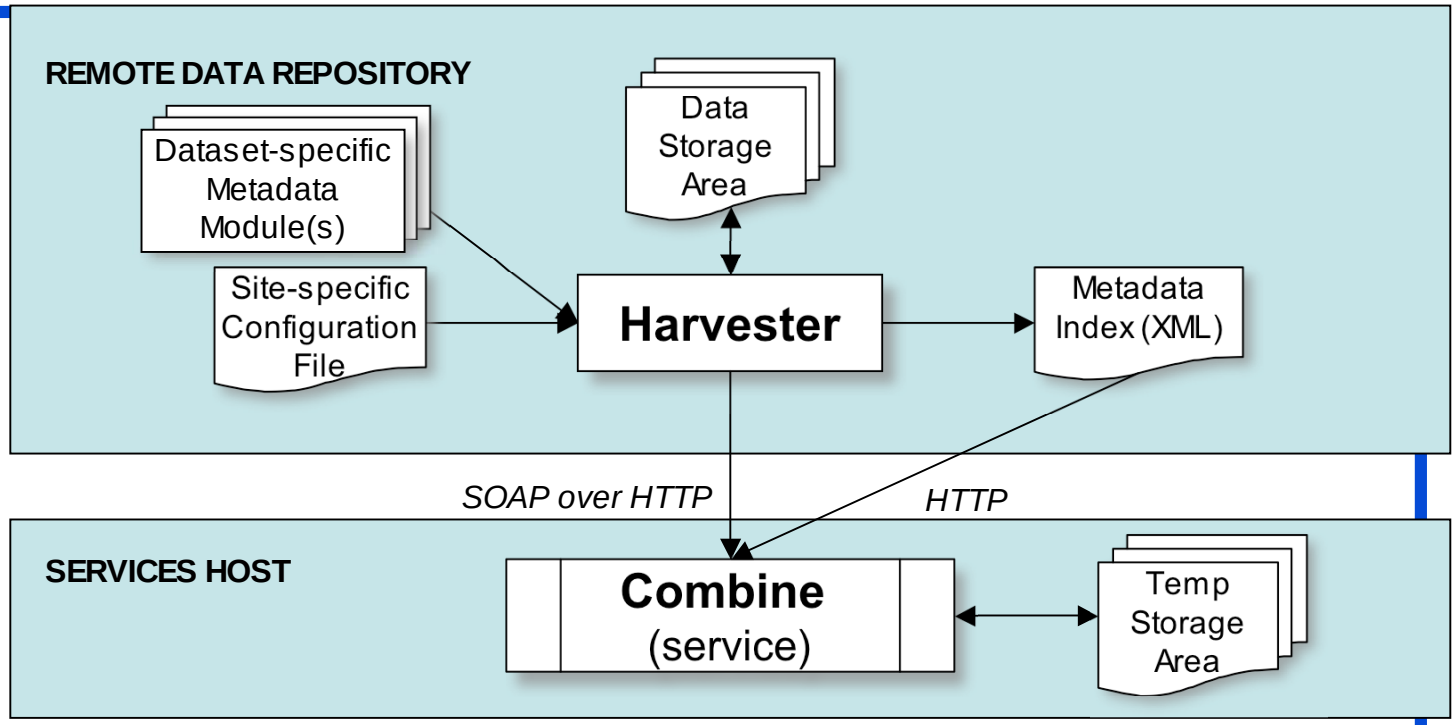
A metadata index of the remote repository is generated and maintained at the site.



Combine

The Harvester notifies a Combine at a services host (via SOAP interfaces), of the availability of the index.

The Combine retrieves the index to a local temporary work area.



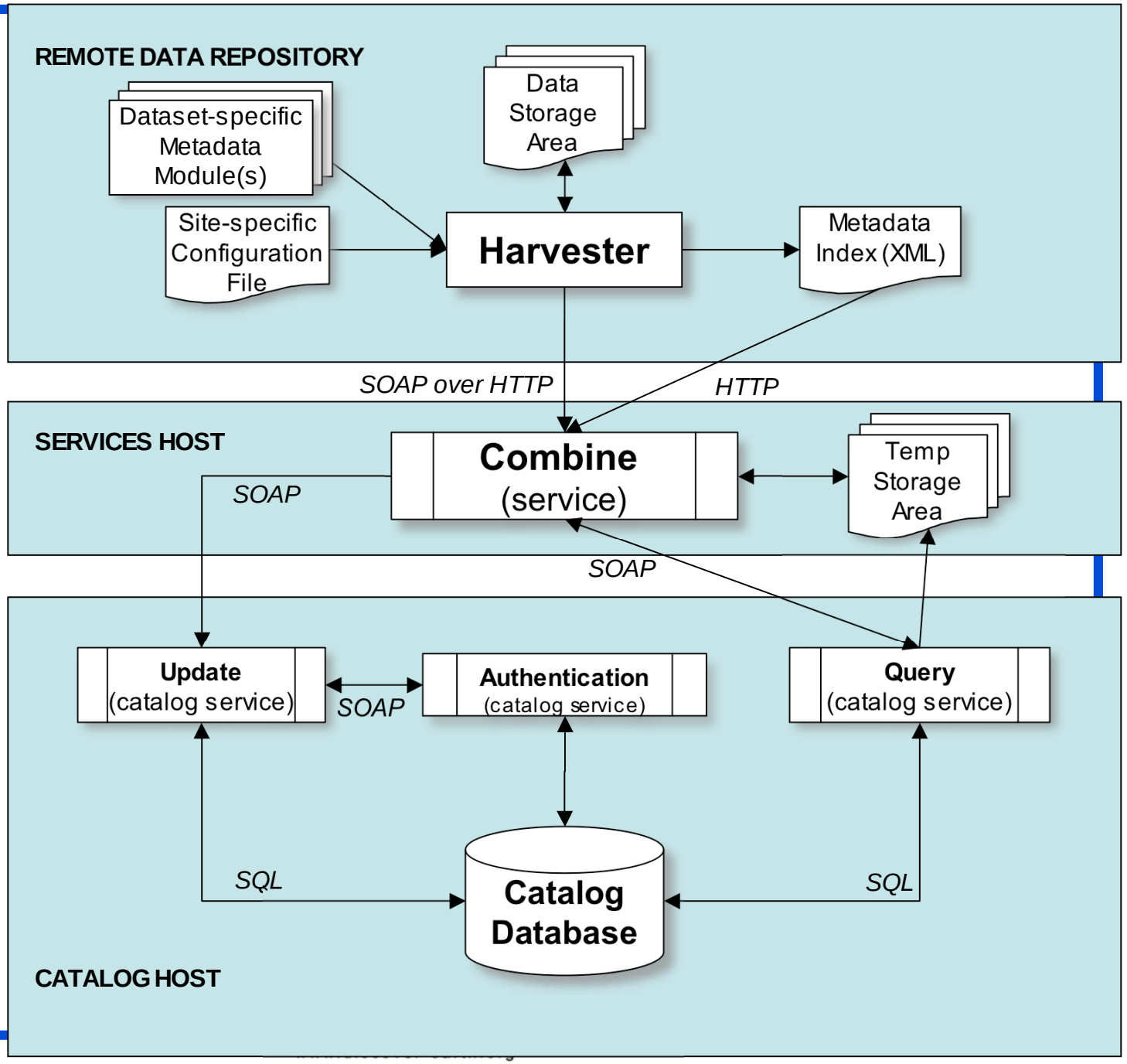
Catalog Services

Catalog services insulate the Combine from database specifics for queries and updates.

Query service returns current inventory information to Combine's temporary storage

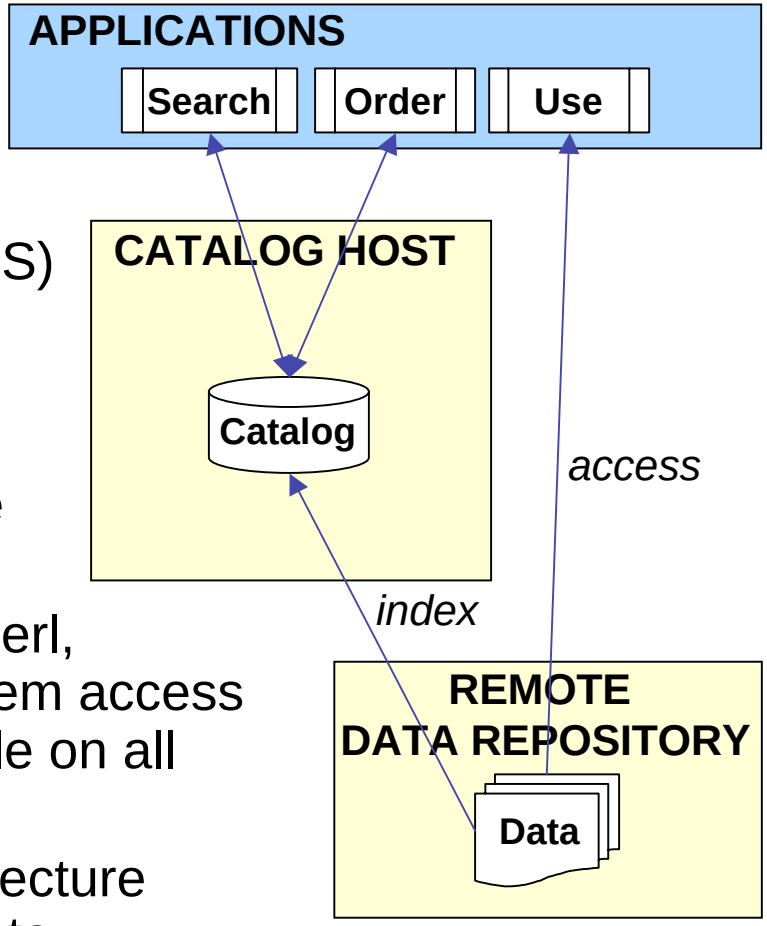
Combine cross checks the new metadata with existing inventory and determines updates needed.

Update service requires authentication before writing to the catalog database.



Conclusions

- The Harvester-Combine approach effectively utilizes a service oriented architecture to harvest inventory information for distributed online repositories (in testing at UAH and RSS)
- A services approach insulates the various distributed components from platform and language dependencies, and from changes to other parts of the system.
- The Harvester-Combine is written in Perl, utilizing platform-independent file system access modules so it has proven to be portable on all tested systems.
- This distributed service oriented architecture should prove to be a scalable solution to indexing remote repositories.



Related / Future Work

Technology research in distributed services

- Service oriented architecture for data management and processing
- Data / service semantics and knowledge management
- Techniques to transfer large binary data between services
- Service discovery and workflow formulation
- Automated service invocation
- Express data delivery services via subscription

Acknowledgements and Links

DISCOVER partners at the National Space Science and Technology Center (NSSTC) and at Remote Sensing Systems have assisted in the deployment and testing of these components.

DISCOVER Project: http://www.discover-earth.org

DISCOVER IT: http://discover.itsc.uah.edu

GHRC Data Pool: http://datapool.nsstc.nasa.gov